

# BAYESIAN SENSITIVITY ANALYSIS WITH FISHER-RAO METRIC

SEBASTIAN KURTEK AND KARTHIK BHARATH

**ABSTRACT.** We propose a geometric framework to assess sensitivity of Bayesian procedures to modeling assumptions based on the nonparametric Fisher-Rao metric on the non-linear manifold of probability densities. While the framework is general in spirit, the focus of this article is restricted to metric-based diagnosis under two settings: assessing local and global robustness in Bayesian procedures to perturbations of the prior; identification of influential observations in a Bayesian regression setting with perturbations to the data. The approach is based on the square-root representation of densities which enables us to compute geodesics and geodesic distances in analytical form, facilitating the definition of naturally calibrated discrepancy measures. The approach proposed in this article ought to be viewed as a first step in the investigation of the exploitation of the geometric structure of the space of probability densities in defining metric-based measures of discrepancies under a Bayesian setting.

**Keywords:** Manifold of probability densities; Fisher-Rao metric;  $\epsilon$ -contamination; Influence analysis.

## 1. INTRODUCTION

**1.1. Motivation.** Investigations into the sensitivity of Bayesian inference to the choices of prior, likelihood and loss function have received considerable attention over the years. In particular, issues of Bayesian prior robustness, both global and local, have been studied extensively and measures evaluating robustness have been proposed; we refer the interested reader to the excellent notes, discussions and references in Insua and Ruggeri [2000] for a detailed account of the foundational, methodological and implementation issues. The ascendant feature in a majority of studies in prior robustness has been the predilection towards the use of measures based on variational properties of posterior functionals like the range over a large class of prior perturbations. Problems noted with this approach include the choice of the scale of range of the posterior functional, size of perturbation and its amenability to functionals which are not of ratio-linear type. Ideally, the posterior distributions ranging over the class of perturbations can themselves be compared in a systematic fashion with due consideration given to the underlying geometry of the space of posterior densities while developing sensitivity measures.

In this article, motivated by the geometric approach to Bayesian influence analysis proposed in the elegant work by Zhu et al. [2011], we investigate the utility in incorporating geometric information into the development of measures to assess Bayes sensitivity. More generally, we develop a metric-based framework for comparing probability densities but restrict our attention to problems in a Bayesian setting. We demonstrate the advantages arising out of incorporating the geometrical information of the underlying space of posterior densities under two settings: assessing local and global robustness in Bayesian procedures to perturbations of the prior via a contamination class (Berger [1994, 1990], Berger and Berliner [1986]); identification of influential observations in a Bayesian regression setting based on perturbations to the data via case-deletions (Dey and Birmiwal [1994], Guttman and Pena [1993], Carlin and Polson [1991]). For the settings aforementioned, we propose sensitivity measures

developed using distances based on an intrinsic metric between posterior distributions which, importantly, possesses a natural calibrated scale. This obviates the need to devise a calibration mechanism which might otherwise have a strong influence on the subsequent inferential procedures. The circumscription of our approach to the two settings is not a limitation as it is not our objective to propose a unified framework for Bayesian sensitivity analysis; rather, we hope to highlight the utility in employing our framework in procedures which are based on comparing densities. Our key motivation is to provide a framework for examining sensitivity to perturbations of the prior and data in a Bayesian setting which can be easily used in practice.

**1.2. Features of our approach, and organization of the article.** In section 2, we consider the Banach manifold of probability density functions and offer a way to calculate geodesic paths and distances between *any* pair of densities analytically. As will be elucidated in the sequel, this will be done by employing the intrinsic nonparametric Fisher-Rao Riemannian metric proposed by Rao [1945] in conjunction with a convenient choice of representation of the densities by the square-root transformation proposed by Bhattacharya [1943]. It will be shown that this transformation takes us from the non-linear manifold of probability densities to the positive orthant of unit Hilbert hypersphere, geometry of which is well understood. Several examples are provided to aid intuition behind our proposed approach.

Sensitivity measures based on comparing the posterior distributions using divergence measures have been considered before; for example, see Dey and Birmiwal [1994]. The chief drawback in using a divergence measure is that it does not satisfy the requirements of a metric—the popular Kullback-Leibler divergence, for instance, does not satisfy the properties of symmetry and triangle inequality. As a consequence, there is a serious issue with calibration of quantities based on divergence measures, or lack thereof. A surreptitious assumption in most procedures employing divergence measures is that the statistical implications of the procedure are not overly affected by the absence of a metric structure in the divergence measure. One obvious reason behind this particular drawback with divergence measures is that their definition is independent of the geometry of the underlying space of distributions; relatedly, the parametric Fisher information Riemannian metric on the manifold of densities can be viewed as the infinitesimal form of the Kullback-Leibler divergence in the sense that the metric is the second derivative of the Kullback-Leibler divergence. Using the parametric version of the Fisher-Information metric, geodesic distances between parametric densities on statistical manifolds have been derived under very special cases; see for instance, p. 235 in Amari et al. [1987]. However, these geodesic distances, in general, are not easy to compute and are usually approximated (Carter et al. [2009], Zhu et al. [2011]).

In the issue of choice of prior perturbations, while several advances have been made in identifying the appropriate perturbation class, geometrical considerations in its construction have received scant attention. To this end, in section 3, we construct a geometric  $\epsilon$ -contamination class of priors which represent geometric perturbations of the base prior where the  $\epsilon$  has an interpretation as a fraction of a distance along geodesics. We employ geodesic distances between posterior densities as a natural measure of discrepancy before and after contamination; indeed, functionals of the posterior can then be evaluated at the ‘nearest’ and the ‘farthest’ to provide a range for the posterior measure of interest. Several illustrative examples are provided comparing geodesic distances with Kullback-Leibler divergences. Using the geodesic distance between probability densities, we furthermore, investigate the local robustness properties of the resulting posteriors or some commonly used quantities like the Bayes factor when the prior is perturbed infinitesimally.

The second issue considered in this article is in the identification of influential cases under a regression setting via perturbations to the data in the form of case-deletion; this comprises section 4. Influence measures based on a general  $\phi$ -divergence (Cziszar [1974]) between posterior densities for identifying outlying response values and influential cases in a Bayesian regression model was considered in Peng and Dey [1995]. We proceed along similar lines and replace their divergence-based measure with the geodesic distance between the posterior densities of interest and examine its properties. Performance of the proposed measures is examined on two datasets, pertaining to linear and logistic regression settings. Comparisons of our results with theirs and the popular Cook's distance are provided and advantage enjoyed by our discrepancy measure because of the natural calibration is clearly demonstrated. Finally, in section 5, we summarize some of the salient features of our approach, discuss some of the shortcomings and comment on possible extensions. Proofs of results are relegated to the Appendix section.

## 2. GEOMETRIC PRELIMINARIES

**2.1. Fisher-Rao Metric.** For simplicity, we shall restrict our attention to the case of univariate densities on  $\mathbb{R}$ . We note, however, that the framework is equally valid for all finite dimensional distributions. Denote by  $\mathcal{P}$ , the Banach manifold of probability density functions, defined as

$$\mathcal{P} = \left\{ p : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0} : \int_{\mathbb{R}} p(x) dx = 1 \right\}.$$

The space  $\mathcal{P}$  is not a vector space but a manifold with a boundary because any density function whose value is zero for any  $x \in \mathbb{R}$  is a boundary element. For any point  $p$  in the interior of  $\mathcal{P}$ , define the tangent space as follows:

$$T_p(\mathcal{P}) = \left\{ \delta p : \mathbb{R} \rightarrow \mathbb{R} : \int_{\mathbb{R}} \delta p(x) p(x) dx = 0 \right\}.$$

Intuitively, the tangent space at any point  $p$  on the manifold  $\mathcal{P}$  contains all possible perturbations of the density function  $p$ . For any two tangent vectors  $\delta p_1, \delta p_2 \in T_p(\mathcal{P})$ , the nonparametric version of the Fisher-Rao Riemannian metric is given by (see Rao [1945], Amari [1985], Vos and Kass [1997])

$$(2.1) \quad \langle \langle \delta p_1, \delta p_2 \rangle \rangle_p = \int_{\mathbb{R}} \delta p_1(x) \delta p_2(x) \frac{1}{p(x)} dx.$$

Under the intrinsic Fisher-Rao metric, it is to be noted that the linear interpolation  $(1 - \alpha)p_1 + \alpha p_2$  for  $0 < \alpha < 1$  is not a geodesic between  $p_1$  and  $p_2$  in  $\mathcal{P}$ . An important property of this metric is that a re-parameterization of densities does not change distances between them. In other words, the action of the reparameterization group is by isometries; see Čencov [1982] for details. The nonparametric version of the Fisher-Rao metric has not been accorded the attention it deserves within the statistical community. One of the goals of this article is in highlighting the utility of this metric as a flexible and easily implementable tool in developing statistical methodologies. This metric has already proven to be very useful for various tasks in computer vision, shape analysis and functional data analysis by Srivastava et al. [2007, 2011], Srivastava et al. [2011]. In those works, the authors note that warping or reparameterization functions can be viewed as cumulative distribution functions on  $[0, 1]$  and their derivatives then become densities.

*Remark 1.* The Fisher-Rao metric defined in (2.1) is closely related to the Fisher information rendering it attractive in statistical methodologies. To elaborate on this relationship, we will

utilize the parametric version of the Fisher-Rao metric. Consider the manifold generated by a parametric family of probability density functions  $\mathcal{F} = \{f(x, \theta) | \theta \in \Theta\}$ . Then, the norm on  $\mathcal{F}$  is induced by the Fisher-Rao Riemannian metric

$$\begin{aligned} \int_{\mathbb{R}} \left( \frac{\partial}{\partial \theta} f(x, \theta) \right)^2 \frac{1}{f(x, \theta)} dx &= \int_{\mathbb{R}} \left( \frac{\partial}{\partial \theta} \log(f(x, \theta)) \right)^2 f(x, \theta) dx \\ &= E_{\theta} \left[ \frac{\partial}{\partial \theta} \log(f(x, \theta)) \right]^2, \end{aligned}$$

which is the Fisher information matrix. Clearly, the metric is also closely related to the Cramer-Rao lower bound on the estimation error of a parameter. This connection between the parametric version of the Fisher-Rao metric and Fisher information provides a compelling motivation for our use of the nonparametric version of this metric.

**2.2. Representation Space of Probability Density Functions.** One major drawback of using the Fisher-Rao metric in practice is the difficulty of computing geodesic paths and distances. This difficulty comes from the fact that the Riemannian metric changes from point to point on the manifold. It is hence required to choose a suitable representation of the space  $\mathcal{P}$  which would facilitate the computation of the geodesic paths and distances; indeed, this would result in a new manifold. Depending on the choice of the representation, the resulting Riemannian structure can have varying degrees of complexity requiring numerical techniques to approximate geodesics. Choice of representations include the CDF, the log density etc.; both these representations do not alleviate the problem of computing geodesics (Srivastava et al. [2007]). The log density representation was employed in Zhu et al. [2011] in their geometric approach to Bayesian influence analysis. The practical implementation of their approach was stymied by the need to approximate the geodesic distances numerically unavailable in closed-form.

The square-root representation proposed by Bhattacharya [1943] proffers a solution to this debilitating issue. Bhattacharya [1943] showed that a square-root transformation of the probability density functions simplifies the Riemannian metric and geometry of the space. In particular, the Fisher-Rao metric becomes the standard  $\mathbb{L}^2$  metric and the space of probability density functions becomes the positive orthant of the unit hypersphere in  $\mathbb{L}^2$ .

**Definition 1.** Define a continuous mapping  $\phi : \mathcal{P} \mapsto \Psi$  where the space  $\Psi$  is the space containing the positive square root of all possible density function. As a consequence, define the square-root transform (SRT) of probability density functions as  $\phi(p) = \psi = +\sqrt{p}$ . Note, that the inverse mapping is simply  $\phi^{-1}(\psi) = p = \psi^2$ .

The space of all square-root transform representations of probability density functions is

$$\Psi = \left\{ \psi : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0} : \int_{\mathbb{R}} |\psi(x)|^2 dx = 1 \right\}$$

and represents the positive orthant of the Hilbert sphere. We shall henceforth omit the  $+$  sign from the representation for notational convenience. The SRT's utility in computing geodesic paths and distances under the Fisher-Rao metric is characterized by the following theorem in Bhattacharya [1943]:

**Theorem 1.** *Under the SRT representation, the Fisher-Rao metric reduces to the standard  $\mathbb{L}^2$  metric.*

*Remark 2.* The square-root representation and the resulting  $\mathbb{L}^2$  metric bears a striking similarity to the Hellinger distance between two probability densities with respect to a common

dominating measure. The Hellinger distance defines a bounded metric on the space of probability densities and generates the same topology as that of the Total Variation distance. However, the Hellinger distance is an *extrinsic* metric defined on the ambient space of the space of probability densities whereas the Fisher-Rao metric is an intrinsic Riemannian metric defined based on the geometry of the space. Our objective of using the geometry underlying the space of densities in defining sensitivity measures, motivates our use of the Fisher-Rao metric.

The theorem encapsulates the advantages associated with the SRT in analyzing probability density functions: first, the complicated Fisher-Rao metric simplifies to the standard  $\mathbb{L}^2$  metric; second, the space of probability density functions becomes the positive orthant of the unit hypersphere. Since the differential geometry of the sphere is well known, one can compute geodesic paths and distances between probability density functions analytically. Our general approach in the remainder of the paper will be to represent probability density functions using their SRT representation, compute geodesic paths and distances on  $\Psi$ , and then map them back to  $\mathcal{P}$  using the inverse mapping provided in Definition 1.

**2.3. Geometry of Unit Hilbert Hypersphere.** Based on the definitions of  $\mathcal{P}$  and  $\Psi$ , for a probability density function  $p \in \mathcal{P}$ ,

$$\int_{\mathbb{R}} |p(x)| dx = \int_{\mathbb{R}} |\psi(x)|^2 dx = 1;$$

in other words, the  $\mathbb{L}^2$  norm of  $\psi$  is 1. The set of all such  $\psi$  is the positive orthant of a unit sphere in  $\mathbb{L}^2$ . It is also a Hilbert manifold (see Lang [1999]). Since  $\mathbb{L}^2$  is a vector space, its tangent space  $T_f(\mathbb{L}^2) = \mathbb{L}^2$  for all  $f \in \mathbb{L}^2$ . The  $\mathbb{L}^2$  Riemannian metric for any  $v_1, v_2 \in \mathbb{L}^2$  is the usual inner product

$$\langle v_1, v_2 \rangle = \int_{\mathbb{R}} v_1(x) v_2(x) dx.$$

A Riemannian structure can be endowed on  $\Psi$  in a straightforward fashion: for a  $\psi \in \Psi$ , the tangent space at  $\psi$  is given by

$$T_\psi(\Psi) = \{v \in \mathbb{L}^2 \mid \langle v, \psi \rangle = 0\};$$

i.e. the vectors tangent to a sphere at a point are orthogonal to the direction representing that point. Once we impose the  $\mathbb{L}^2$  metric on the tangent spaces of  $\Psi$ , we are rewarded with a Riemannian structure. The fact that  $\Psi$  is the positive orthant of a sphere is important because the differential geometry of a sphere is well understood.

For statistical purposes, we are interested in the geodesic path and distance between two points in  $\Psi$ . Observe that since we are on the unit infinite dimensional sphere, the geodesic distance between any two points or vectors (densities) is given by the angle between them; in other words, if  $\psi_1$  and  $\psi_2$  are two points in  $\Psi$ , based on the nonparametric Fisher-Rao metric, the *geodesic distance* between them is given by

$$\theta = \cos^{-1}(\langle \psi_1, \psi_2 \rangle).$$

The geodesic path between  $\psi_1$  and  $\psi_2$  will be indexed by  $\tau \in [0, 1]$  and the length of the path is the geodesic distance. Formally, for any  $\psi_1, \psi_2 \in \Psi$ , the *geodesic path* between them in  $\Psi$ , for  $\tau \in [0, 1]$ , is given by

$$(2.2) \quad \eta_\tau(\psi_1, \psi_2) = \frac{1}{\sin(\theta)} [\sin(\theta - \tau\theta)\psi_1 + \sin(\tau\theta)\psi_2], \quad \theta = \cos^{-1}(\langle \psi_1, \psi_2 \rangle).$$

By varying  $\tau$  in  $[0, 1]$  one traverses the path between  $\psi_1$  and  $\psi_2$ . The restriction to the positive orthant of the unit sphere does not pose any additional difficulties; for two points  $\psi_1, \psi_2 \in \Psi$  the shortest geodesic between them is entirely contained in  $\Psi$ . It is easy to see that  $\theta$  is bounded above by  $\pi/2$ ; in other words, the farthest two densities can be is  $\pi/2$ . *This imposes an upper bound on the geodesic distance between probability densities leading to a naturally calibrated discrepancy measure which can be used directly while comparing densities.* We will exploit this representation and the ensuing geometrical properties for performing subsequent tasks. It is easy to note the ramifications of a calibrated metric on the space of densities on statistical procedures: Not only are we handed a proper metric to compare probability densities, it also is the case that this metric is bounded above by  $\pi/2$ . It is conceivable at this point, that a myriad of potential statistical procedures, based on the geodesic distance between densities, can be developed; we note such a possibility, but in the interests of brevity, restrict our attention to assessing robustness in a Bayesian setting to perturbations of the prior and the data.

On occasions in statistical applications, one is interested in discovering target densities which are at a specified distance along a particular direction from a given  $\psi_1$ . For example, the construction of an  $\epsilon$ -contamination class for checking prior robustness is based along this line of reasoning albeit, admittedly,  $\epsilon$  is not really a “distance” in the conventional sense. For such purposes, a geodesic on  $\Psi$  can also be characterized in terms of a direction given by a tangent vector  $v \in T_{\psi_1}(\Psi)$  at a distance given by the length of  $v$  from  $\psi$ , as

$$(2.3) \quad \eta_\tau(\psi_1, v) = \cos(\tau\|v\|)\psi_1 + \sin(\tau\|v\|)\frac{v}{\|v\|}.$$

The two representations of the geodesic paths given in equations (2.2) and (2.3) are equivalent in the sense that the first one, upon fixing two points  $\psi_1$  and  $\psi_2$  in  $\Psi$ , provides the shortest path between them; equation (2.3) upon fixing, say  $\psi_1$  and a direction  $v$  towards  $\psi_2$ , in the tangent space of  $\psi_1$ , represents the equation of the geodesic path from  $\psi_1$  along  $v$  at a distance  $\|v\|$ . It is clear now that in order to calculate geodesic paths and distances between two points in  $\Psi$ , one needs to move freely between the tangent spaces at the points and  $\Psi$ . In order to facilitate this mechanism, we define the exponential map at a point  $\psi_1 \in \Psi$ , denoted by  $\exp : T_{\psi_1}(\Psi) \mapsto \Psi$ , as

$$\exp_{\psi_1}(v) = \cos(\|v\|)\psi_1 + \sin(\|v\|)\frac{v}{\|v\|}.$$

The purpose of this map is to map points from the tangent space at a point  $\psi_1$  to the space  $\Psi$ . In the other direction, the inverse of the exponential map, denoted by  $\exp_{\psi_1}^{-1} : \Psi \mapsto T_{\psi_1}(\Psi)$ , is given by

$$\exp_{\psi_1}^{-1}(\psi_2) = \left[ \frac{\theta}{\sin(\theta)} (\psi_2 - \cos(\theta)\psi_1) \right],$$

and can be used to map points from the representation space of probability density functions to the tangent space at a point  $\psi_1$ . The preceding discussion, the respective quantities and their relationships are nicely encapsulated in the Figure 1.

**2.4. Illustrative examples.** We consider four examples wherein geodesic paths and distances are computed between some common densities; the last one is a bivariate density example emphasizing the generality of this approach to finite dimensional densities as well. Figure 2 summarizes the results from the first three examples. We compare our approach with the straight line distance (technically, a linear interpolation) between the densities which contains no geometric information of the underlying space. Densities which are present along the

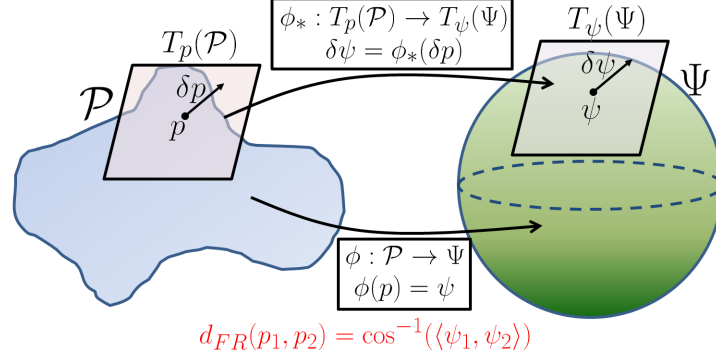


FIGURE 1. The Figure provides a succinct description of the square-root transformation from  $\mathcal{P}$  to the unit Hilbert sphere  $\Psi$ . On  $\mathcal{P}$ , at a point (density)  $p$ , its tangent space  $T_p(\mathcal{P})$  is shown along with the corresponding tangent vector  $\delta p$ . These quantities are mapped to the tangent space of  $\psi$  on  $\Psi$  and the counterparts are displayed in a similar manner. The key point being the isometric property of the square-root representation with the distance  $d_{FR}(p_1, p_2)$  between  $p_1$  and  $p_2$  being exactly equal to the distance between its mappings on  $\Psi$  given by arc cosine of the angle between  $\psi_1$  and  $\psi_2$ .

straight line path and the geodesic path between the two points are plotted and compared to the density at the midpoint to highlight the difference. In addition, we contrast the values of the geodesic distance  $d_{FR}$  with the Kullback-Leibler divergence (KL).

**Example 1.** We consider the two normal densities  $p_1 \sim N(-2, 1)$  and  $p_2 \sim N(2, 1)$ . In the first row of Figure 2, denoted by (a), note that the Fisher-Rao geodesic path is quite different from the straight line path between these densities. This difference is highlighted by plotting the midpoint of each path in the third column. Based on the Fisher-Rao distance of 1.435, we note that  $p_1$  and  $p_2$  are far apart since the upper bound is  $\pi/2 \approx 1.507$ . In this simple example, the Kullback-Leibler divergence can be computed analytically and turns out to be symmetric; this is so owing to the fact that the scale parameters of both densities are the same. Nonetheless, it is difficult to ascertain from this number whether the two densities are considered close or far under the KL divergence since the KL divergence does not possess a natural calibration.

**Example 2.** Here  $p_1 \sim N(0, 1)$  and  $p_2 \sim t_1$ , a student's  $t$  distribution with one degree of freedom (The row corresponding to (b) in Figure 2). The Fisher-Rao distance between the standard normal and the  $t$  with the heaviest tail is 0.47. Note that this number is much smaller compared to two normals with differing location parameters in Example 1. We are in a position to directly compare the two numbers owing to the natural calibration. The most striking result here is the high disparity between the KL divergences with the arguments interchanged;  $KL(p_1, p_2)$  is 0.25 whereas  $KL(p_2, p_1)$  is 3.308 emphasizing the problems associated with lack of symmetry. One could try to symmetrize the KL divergence by using  $(KL(p_1, p_2) + KL(p_2, p_1))/2$ , which in this case would yield 1.7840, and is, nevertheless, hard to interpret owing to the absence of calibration.

**Example 3.** In row (c) of Figure 2,  $p_1 \sim N(0, 1)$  and  $p_2 \sim SN(5)$ , where  $SN(5)$  denotes a skew-normal density with skewness parameter 5. Again, notice the disconcerting disparity in

the KL divergences obtained upon switching the arguments. the Fisher-Rao distance is 0.67 in this case.

**Example 4.** In this example, presented in Figure 2, we consider two bivariate normal distributions  $p_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$  and  $p_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$  where

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0.5 \\ 0.2 \end{bmatrix}, \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.2 & 0.4 \\ 0.4 & 0.6 \end{bmatrix} \text{ and } \boldsymbol{\mu}_2 = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.5 & -0.2 \\ -0.2 & 0.7 \end{bmatrix}.$$

The Kullback Leibler divergence is close to being symmetric in this case. The Fisher Rao distance is 0.7151.

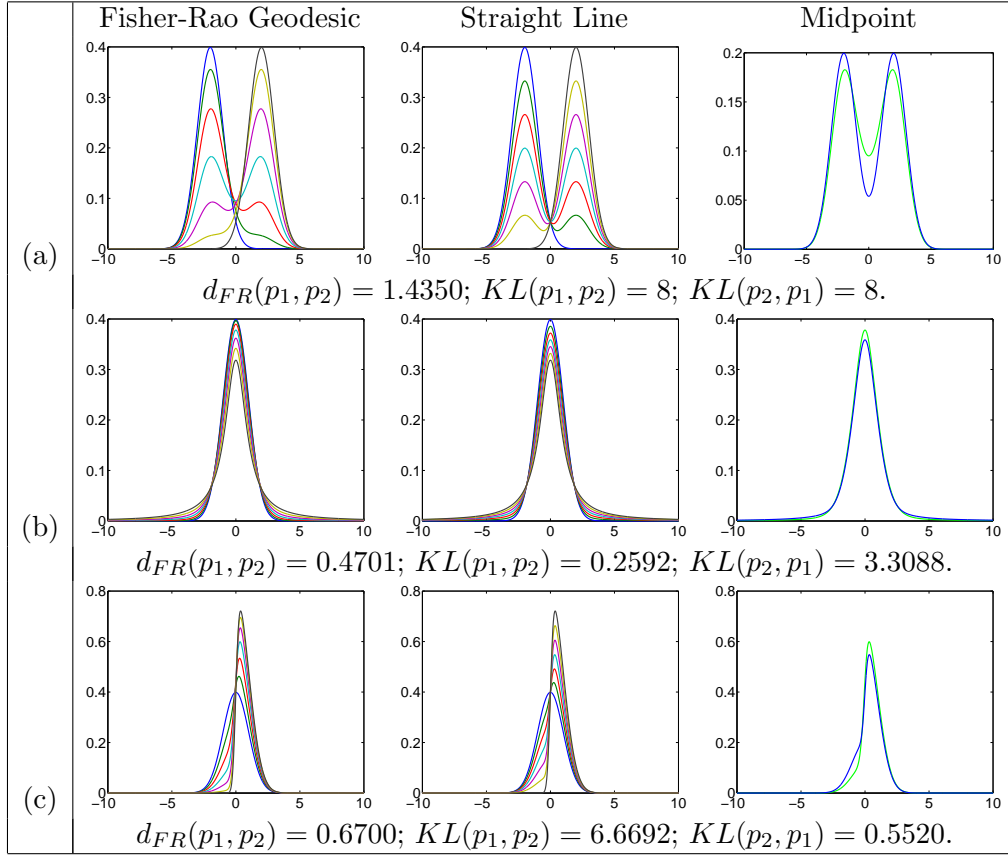


FIGURE 2. Geodesic paths and distances between different distributions. We compare the Fisher-Rao geodesic path and a straight line path between two densities  $p_1$  (blue) and  $p_2$  (black). This path is sampled at 7 equally spaced points. The third column displays the midpoint of the FR path (green) and the straight line path (blue). We also compare the Fisher-Rao geodesic distance with the Kullback Leibler divergence. (a)  $N(-2,1)$  and  $N(2,1)$ , (b)  $N(0,1)$  and  $t_1$ , (c)  $N(0,1)$  and  $SN(5)$ .

### 3. PRIOR ROBUSTNESS

We are now in a position to employ the ingredients described in the preceding section to develop a suitable geometric framework for assessing prior robustness. Our methodology



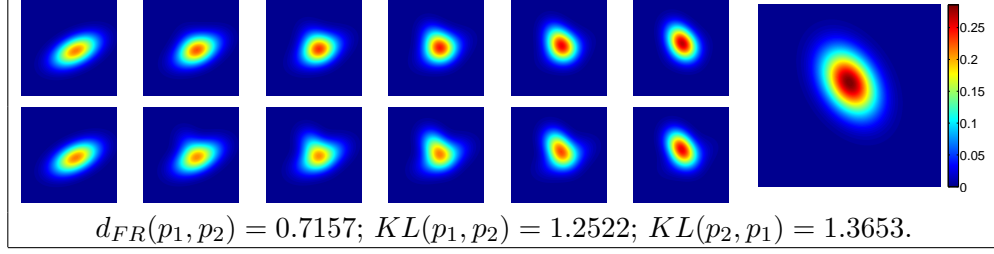


FIGURE 3. Geodesic path and distance between two bivariate normal distributions. We again compare the Fisher-Rao geodesic path (top) with a straight line interpolation (bottom). Note that the endpoint (and starting point) in both cases is the same. We also compare the Fisher-Rao distance with the Kullback Leibler divergence.

is centered around developing robustness measures by means of geodesic distances between posterior densities. A note about the notation used in this section:  $X$  denotes the observable random variable which will be assumed to have a density  $f(x|\theta)$  with respect to the Lebesgue measure where  $\theta$  is a vector (finite or infinite) of unknown parameters lying in a parameter space  $\Theta$ . A prior density on  $\Theta$  is denoted by  $\pi$  and the resulting posterior distribution of  $\theta$  obtained by the Bayes rule, assuming it exists, will be denoted by  $p_\pi(\cdot|x)$  and is defined by

$$p_\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x|\pi)},$$

where  $m(x|\pi)$  is the marginal density of  $X$  obtained by averaging over the prior  $\pi$  given by

$$m(x|\pi) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta.$$

**3.1. Geometric  $\epsilon$ -contamination class.** Let  $\pi_0$  represent a baseline prior probability density on the parameter  $\theta$  and  $\mathcal{G}$  denote the class of contaminants or perturbations which are bonafide densities. This class  $\mathcal{G}$  is appropriately constructed based on the problem of interest and the baseline prior. In current literature, the set  $\Gamma$  of  $\epsilon$ -contaminated priors is most commonly defined as

$$(3.1) \quad \Gamma = \left\{ (1 - \epsilon)\pi + \epsilon g; 0 \leq \epsilon \leq 1, g \in \mathcal{G} \right\}.$$

We will henceforth refer to this class as the linear  $\epsilon$ -contamination class. Owing to its structure it is easy to see that the class  $\Gamma$  induces a similar kind of contamination on the marginal  $m$  as

$$\Gamma_m = \left\{ (1 - \epsilon)m(x|\pi) + \epsilon m(x|g); 0 \leq \epsilon \leq 1, g \in \mathcal{G} \right\},$$

and the ensuing class of contaminated posteriors corresponding to  $\Gamma$  can be written as

$$\Gamma_p = \left\{ \lambda(x)p_\pi(\theta|x) + (1 - \lambda(x))p_g(\theta|x); 0 \leq \epsilon \leq 1, g \in \mathcal{G} \right\},$$

where  $\lambda(x) = (1 - \epsilon)m(x|\pi) [(1 - \epsilon)m(x|\pi) + \epsilon m(x|g)]^{-1}$ . Note how the linear nature of the perturbations of characterizing the class  $\Gamma$ , in a certain sense, permeates through to the marginal and the posterior. One interpretation of  $\epsilon$  is as a measure of uncertainty regarding the choice of the original prior  $\pi$  (Moreno [2000]). If one were to adopt this interpretation,

then under the linear  $\epsilon$ -contamination class, the amount of uncertainty regarding  $\pi$  carries over *exactly* while expressing the amount of uncertainty regarding the marginal which is averaged over the prior. Indeed, if one is uncertain regarding the choice of the likelihood too, then such a phenomenon is quite undesirable. Nevertheless, the contamination class  $\Gamma$  has a nice interpretation in terms of mixtures of densities; but it disregards the underlying geometry of the space of probability density functions.

We wish to tie the  $\epsilon$  with actual distances on the perturbation space. In other words, it makes sense in our setting, loosely speaking, to *perturb  $\pi$  by moving away from it in the direction  $v_g$  by a small distance  $\epsilon\|v\|$* . Since  $\epsilon \in [0, 1]$ ,  $\epsilon\|v\|$  represents the fraction of  $\|v_g\|$  along the direction  $v_g$ . We can utilize the SRT representation to generate a geometric notion of the set of  $\epsilon$ -contaminated priors. This would then represent a true perturbation adhering to the geometry of the space, and would retain the mixture interpretation, albeit under the Fisher-Rao metric.

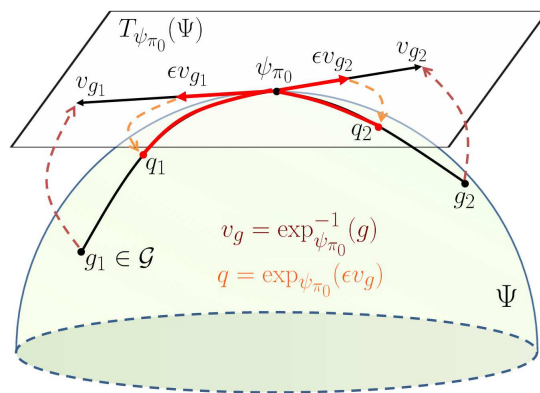
Before defining the geometric  $\epsilon$ -contamination class we note that our interest is to create a framework which can easily be used in practice. It is commonly the case that one is unsure about the prior and wishes to verify the sensitivity of statistical conclusions to a finite set of prior alternatives. For example in a regression setting, a common prior on the regression coefficients is a multivariate normal; one might be interested in exploring the effect of having a multivariate  $t$  prior with heavier tails at a finite number of the degrees of freedom parameter; or one might wish to check for the sensitivity against a skew-normal class indexed by the skewness parameters at a finite set of values. Taking into account such considerations, we restrict our attention to a finite class of perturbations in our examples and illustrations noting that in principle, the class can be infinite. Let  $\psi_{\pi_0}$  be the SRT representation of the baseline prior  $\pi_0$  and  $\psi_{g_1}, \dots, \psi_{g_n}$  be the SRT representations of all elements of a finite class  $\mathcal{G} = \{g_1, \dots, g_n\}$  of contaminants. One can construct a set of tangent vectors  $v_{g_1}, \dots, v_{g_n} \in T_{\psi_{\pi_0}}(\Psi)$  using the inverse exponential map as  $v_{g_i} = \exp_{\psi_{\pi_0}}^{-1}(\psi_{g_i})$ ,  $i = 1, \dots, n$ . This provides us with a finite set of different perturbations of the baseline prior.

**Definition 2.** For a class of densities  $\mathcal{G} = \{g_1, \dots, g_n\}$ , the geometric  $\epsilon$ -contamination class corresponding to the prior  $\pi_0$  is defined as

$$(3.2) \quad \tilde{\Gamma} = \left\{ \phi^{-1}(\exp_{\psi_{\pi_0}}(\epsilon v_{g_i})); 0 \leq \epsilon \leq 1, g_i \in \mathcal{G}, 1 \leq i \leq n \right\}.$$

Recall that  $\phi^{-1}$  is essentially the operation of squaring the SRT representation. The interpretation of this set is as follows: if one fixes an element  $g_i \in \mathcal{G}$ , then by varying  $\epsilon$  from 0 to 1, one traces the geodesic path from  $\pi_0$  to  $g_i$ . Thus, if we fix a value for  $\epsilon$ , we will obtain a finite set of priors that were contaminated in the direction of  $g_1, \dots, g_n$ . This is described in Figure 4.

*Remark 3.* Some comments are required at this point. The linear contamination class  $\Gamma$  can be rewritten as  $\Gamma = \{\pi + \epsilon(g - \pi); 0 \leq \epsilon \leq 1, g \in \mathcal{G}\}$ . One might be tempted to interpret each element of  $\Gamma$  in the terminology of linear spaces: a small perturbation of magnitude  $\epsilon\|g - \pi\|$  *along* the direction  $(g - \pi)$ . However,  $\Gamma$  is not a linear space in general; convexity can be imposed to ensure that the elements of  $\Gamma$  stay within the class. It is hence inappropriate to view  $\Gamma$  as a true perturbation class in the geometrical sense. In other words, it would not be accurate to view the elements of  $\Gamma$  as densities which are ' $\epsilon\|g - \pi\|$  away from  $\pi$  towards the class  $\mathcal{G}$ '. However, such an interpretation is sensible under our geometrical setting: Given a class  $\mathcal{G}$  or a direction  $v_g$ , we can travel an  $\epsilon$  proportion of the norm of  $v$  along the geodesic path between  $\pi$  and  $g$ .



It is pertinent to note that under geometrical perturbation of the prior, the interpretation of  $\epsilon$  does not carry over to the marginals and the posteriors over the class as in the case of the linear contamination class; the perturbation is non-linear and in principle, it would be difficult to gauge the amount (distances) by which the marginals and the posterior would be perturbed owing to the original perturbation of the prior. However, it is the case that the Fisher-Rao metric satisfies an intuitive but fundamental property regarding the effect of perturbation of the prior on the sampling distribution or the likelihood; any perturbation of the prior ought not to have an effect on the sampling distribution or the data since it represents an infinitesimal change in just the prior component of the model which encompasses other components. As a consequence, we would expect the geometric contamination class based on the nonparametric Fisher-Rao metric to satisfy such a property. The following theorem sheds light on the matter.

**Theorem 2.** *The Fisher-Rao Riemannian metric on the space of joint densities is independent of the sampling distribution under the prior perturbation classes  $\Gamma$  and  $\tilde{\Gamma}$ .*

**3.2. Sensitivity measure and some examples.** Given a likelihood function  $f(x|\theta)$  one can define the baseline posterior density, when it exists, as

$$p_0(\theta|x) = \frac{f(x|\theta)\pi_0(\theta)}{m(x|\pi_0)}.$$

If it is the case that the posterior is given within the constant represented by the integral in the denominator of this expression, we can evaluate it using a numerical integral or by Monte Carlo methods. Note that under the SRT representation, this operation is the same as a simple straight-line projection from  $\mathbb{L}^2$  to  $\Psi$ . In order to compute distances between posterior probability density functions we will again utilize the space  $\Psi$ . We are now given  $p_0(\cdot|x)$ , the baseline posterior, and  $p_{g_1}(\cdot|x), \dots, p_{g_n}(\cdot|x)$ , the set of posteriors generated from the  $\epsilon$ -contaminated priors.

**Definition 3.** For a class of contamination densities  $\mathcal{G}$  consider the geometric  $\epsilon$ -contamination class defined in 2. Then, a measure of sensitivity with respect to the geometric perturbation of  $\pi_0$  is defined as

$$S(\epsilon, \pi_0, \mathcal{G}) = \max \{d_{FR}(p_0, p_{g_i}; g_i \in \mathcal{G}, 1 \leq i \leq n)\},$$

where  $p_{g_i}$  is the perturbed posterior density. Guided by the measure  $S(\epsilon, \pi_0, \mathcal{G})$ , we can additionally compute posterior functionals with respect to the ‘nearest’ and ‘farthest’ posteriors. Note that using the geodesic distance as a measure of robustness in our framework is meaningful because all of the distances are bounded above by  $\pi/2$ . Thus, there is a clear notion of a small distance versus a large distance representing different magnitudes of robustness. Furthermore, such distances capture the geometry of the space of probability density functions.

We consider a couple of examples that showcase the effectiveness of the proposed framework in the context of Bayesian robustness to prior contamination.

**Example 5.** Consider the following simple model:

$$\begin{aligned} x|\theta &\sim f = N(\theta, 1) \\ \theta &\sim \pi_0 = N(0, 1) \end{aligned}$$

We consider a skew normal contamination class, parameterized by a shape parameter  $\alpha \in [-5, 5]$ . In Figure 5, we display the considered  $\epsilon$ -contaminated prior set under the linear (Equation 3.1) and geometric (Equation 3.2) frameworks by fixing  $\epsilon = 0.5$  and  $\alpha = -5, -4, -3, -2, -1, 1, 2, 3, 4, 5$ . Notice that the two methods of contamination appear fairly similar; upon observing the data however, the updated posteriors are shown to be markedly different. We begin by simulating data  $x_1, \dots, x_{50}$  from the baseline model and generating a set of contaminated priors for 31 equally spaced values  $\epsilon \in [0, 1]$  and 101 equally spaced values  $\alpha \in [-5, 5]$  using the two different types of contamination methods. We can compute the baseline posterior density using  $p_0$ , where the normalizing constant is calculated numerically. In a similar fashion, we can compute the posterior density resulting from any of the contaminated priors; we will refer to them as  $p$ . In this example, we utilize the geometric contamination framework for our method (the column (a) in Figure 6) and the linear contamination framework when we evaluate the KL divergence as a robustness measure (columns (b) and (c) in Figure 6). We then compute three distance measures between the baseline posterior density and all of the contaminated posteriors: the Fisher-Rao distance based on the SRT representation of densities, the Kullback Leibler divergence where the expectation is computed with respect to  $p_0$ , and the Kullback Leibler divergence where the expectation is computed with respect to  $p$ . We also compute the posterior mean for  $\epsilon = 0.5$  based on the same set of contaminated models. Note that when  $\alpha = 0$ , the contaminated posterior is the same as the baseline posterior. This procedure is performed on three simulated datasets corresponding to the three rows in Figure 6.

We make a few key observations about the results presented in Figure 6: First, as expected, the KL divergence is asymmetric in all cases, which makes its use as a robustness measure suspect; second, in all cases, the FR distance suggests that the posterior is fairly robust to geometric contamination of the Gaussian prior using skew normal distributions, especially if one takes  $\epsilon$  to be small. The only time this distance becomes relatively high is when  $\epsilon$  is taken to be nearly one and  $\alpha$  is large. Note that when  $\epsilon$  is equal to one, the baseline Gaussian prior is entirely replaced with the skew normal. It can also be seen that the FR distance is more sensitive to departures from  $N(0,1)$  than the KL divergence; the KL divergence appears to pick up departures only for  $\epsilon$  greater than 0.5. We also notice an interesting and important

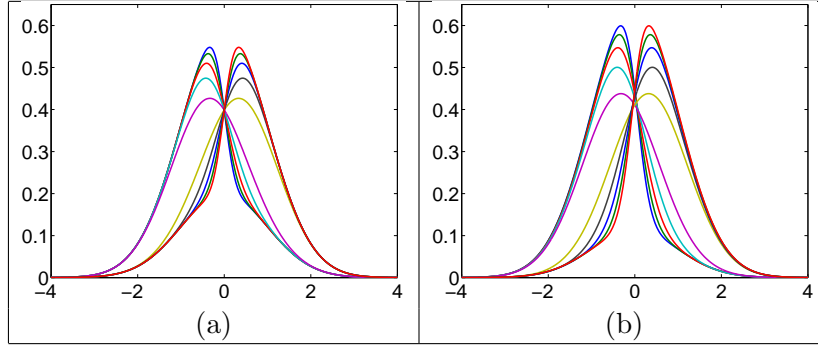


FIGURE 5. Normal prior contaminated using the skew normal contamination class under the (a) linear and (b) geometric frameworks. Observe how the tails are more separated under the geometric framework than in the linear one illustrating the difference between the two methods.

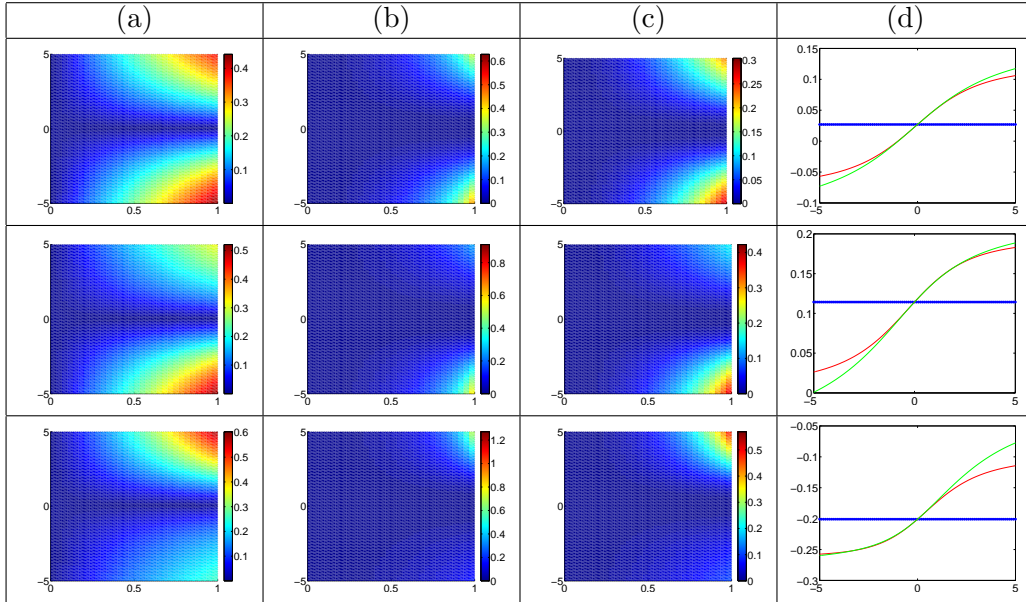


FIGURE 6. Assessment of Bayesian prior robustness for  $\epsilon$ -contamination of a Gaussian prior with a skew-normal distribution. (a) Image of FR distances between baseline and geometrically contaminated posteriors for different values of  $\epsilon$  and  $\alpha$  for 3 simulated datasets (b) Image of KL divergences (expectation computed with respect to  $p_0$ ) between baseline and linearly contaminated posteriors for different values of  $\epsilon$  and  $\alpha$ . (c) Same as (b) but expectation was computed with respect to  $p$ . (d) Posterior means for varying values of  $\alpha$ , where blue is the true posterior mean, green is the posterior mean under geometric contamination and red is the posterior mean under linear contamination.

result from panel (d). The geometric contamination class portrays a more severe departure from the baseline model (in terms of the posterior mean) than the linear contamination class.

This allows for easier identification of severe prior contaminations. This phenomenon is due to the non-linear structure of the geometric contamination class and is consistent with intuition.

**Example 6.** Next, we consider the following baseline model:

$$\begin{aligned} x|\theta &\sim f = N(\theta, 1) \\ \theta &\sim \pi_0 = N(0, 0.2), \end{aligned}$$

and the following data generating model:

$$\begin{aligned} x|\theta &\sim f = N(\theta, 1) \\ \theta &\sim \pi_0 = t_1. \end{aligned}$$

We consider a Student's t contamination class, parameterized by the degrees of freedom  $df$ . As in the previous example, we begin by simulating data  $x_1, \dots, x_{10}$  from the data generating model, and generating a set of contaminated priors for 31 equally spaced values  $\epsilon \in [0, 1]$  and  $df = 3, \dots, 15$ . We present the results of our analysis for one simulated dataset in Figure 7. The shown posterior mean and quantiles were computed for  $\epsilon = 0.5$ .

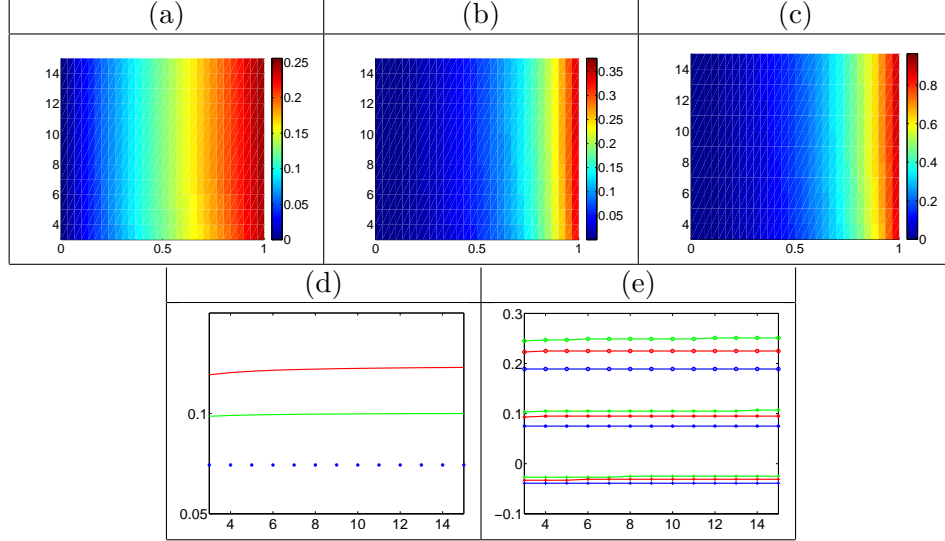


FIGURE 7. Assessment of Bayesian prior robustness for  $\epsilon$ -contamination of a Gaussian prior with a Student's t distribution. (a) Image of FR distances between baseline and geometrically contaminated posteriors for different values of  $\epsilon$  and  $df$ . (b) Image of KL divergences (expectation computed with respect to  $p_0$ ) between baseline and linearly contaminated posteriors for different values of  $\epsilon$  and  $df$ . (c) Same as (b) but expectation was computed with respect to  $p$ . (d) Posterior means for varying values of  $df$ , where blue is the true posterior mean, green is the posterior mean under geometric contamination and red is the posterior mean under linear contamination. (e) Posterior quantiles for varying values of  $df$ , where blue represents the true posterior quantiles, green represents the posterior quantiles under geometric contamination and red represents the posterior quantiles under linear contamination.

Again, we note that the Kullback-Leibler divergence is highly asymmetric in this example. Furthermore, it appears that the Fisher-Rao distance is more conservative. That is, it provides

a value indicating less robustness for lower values of  $\epsilon$  than the Kullback Leibler divergence. As expected, because the data was generated using a heavy-tailed prior, the Fisher-Rao distance increases with increasing values of  $df$  for all  $\epsilon$ . Another interesting result is that the geometric contamination class affects the posterior mean less than the linear contamination class for the same value of  $\epsilon$ . On the other hand, geometric contamination has a greater effect on the posterior quantiles than linear contamination.

**3.3. Local sensitivity to prior perturbation.** In this section we examine first-order local sensitivity to the proposed geometric prior perturbation for the commonly used Bayes factor and a general posterior functional represented via an integral. We then propose a second order local sensitivity measure to assess the effect of geometric  $\epsilon$  prior perturbation on the proposed geodesic distance between posterior densities. As previously stated, let  $\pi_0$  be the baseline prior,  $f$  be the likelihood,  $v_g = \exp_{\psi_{\pi_0}}^{-1}(\psi_g) \in T_{\psi_{\pi_0}}(\Psi)$  be a perturbation vector in the direction of a prior contaminant  $g$  and  $p_0$  be the baseline posterior. Let  $m(x|\pi_0)$  denote the marginal with respect to the baseline prior  $\pi_0$  and define

$$m(x|\epsilon g) = \int_{\Theta} f(x|\theta) \exp_{\sqrt{\pi_0}}(\epsilon v_g)(\theta)^2 d\theta \quad \text{and} \quad m(x|v_g) = \int_{\Theta} f(x|\theta) \sqrt{\pi_0(\theta)} v_g(\theta) d\theta.$$

We shall denote by  $F$  a general functional of interest and denote by  $p_{\epsilon g}$  the posterior obtained from a member of the geometric  $\epsilon$ -contamination class.

**Proposition 1.** *If  $F_{\pi_0}(v_g) = \frac{m(x|\pi_0)}{m(x|\epsilon g)}$  denotes the Bayes factor for comparing the marginals before and after contamination, then*

$$dF_{\pi_0}(v_g) \Big|_{\epsilon=0} = -2 \frac{m(x|v_g)}{m(x|\pi_0)}.$$

**Proposition 2.** *Suppose  $F_{\pi_0,h}(v_g)$  is expectation of  $h(\theta)$  with respect to  $p_{\epsilon g}$ . Then,*

$$dF_{\pi_0,h}(v_g) \Big|_{\epsilon=0} = 2 \int_{\Theta} h(\theta) \frac{f(x|\theta) \sqrt{\pi_0(\theta)} v(\theta)}{m(x|\pi_0)} d\theta - 2 \frac{m(x|v_g)}{m(x|\pi_0)} \int_{\Theta} h(\theta) p_0(\theta|x) d\theta.$$

We now turn our attention to the second order analysis of the squared geodesic distance itself since we employ it as a local influence measure.

**Theorem 3.** *Let  $F_{\pi_0}(v_g)$  represent the geodesic distance between the posteriors  $p_0$  and  $p_{\epsilon g}$ . Then*

$$d^2 F_{\pi_0}(v_g) \Big|_{\epsilon=0} = 4 \frac{m(x|v_g)}{m(x|\pi_0)} \int_{\Theta} \frac{v_g(\theta)}{\sqrt{\pi_0(\theta)}} p_0(\theta|x) d\theta - 2 \int_{\Theta} \frac{v_g(\theta)^2}{\pi_0(\theta)} p_0(\theta|x) d\theta - 2 \frac{m(x|v_g)^2}{m(x|\pi_0)^2}.$$

#### 4. PERTURBATION OF THE DATA

We now turn our attention to sensitivity analysis based on perturbations to the data; we will restrict our attention to a regression setting. The general methodology remains unchanged from the preceding sections in the sense that we will ultimately be developing influence measures based on distances between posteriors. Therefore, we shall illustrate the effectiveness of our approach by considering two datasets: one analyzed under a linear regression setting and the other in a logistic regression setting. The form of perturbation considered in this sections will be case-deletion. Other forms of perturbations can also be analyzed under our framework since our measures are based on the distances between posterior densities.

We fix notation as follows: Let the baseline posterior be  $p_0(\theta|\mathbf{y}, \mathbf{X}) \propto f(\mathbf{y}|\theta, \mathbf{X})\pi(\theta)$ , where  $\theta$  is a vector of unknown coefficients,  $\mathbf{y}$  is the response variable and  $\mathbf{X}$  is the standard design matrix. One can evaluate the influence of the  $k$ th observation on the posterior distribution of  $\theta$  by removing it from the observation set and estimating the posterior distribution using the remaining observations. This results in a new posterior distribution  $p_k(\theta|\mathbf{y}, \mathbf{X}) \propto f(\mathbf{y}_{(k)}|\theta, \mathbf{X}_{(k)})\pi(\theta)$ , where  $\mathbf{y}_{(k)}$  denotes the set of all observations excluding the  $k$ th one.

**Definition 4.** Given the baseline posterior  $p_0$  and the posterior under case deletion  $p_k$ , the influence of observation  $k$  is defined as

$$I(k) = d_{FR}(p_k, p).$$

Note that this distance is symmetric and has an upper bound of  $\pi/2$ , which avoids the ambiguity present in different divergence measures and provides a natural scale for evaluating influence in regression.

**4.1. Linear Regression.** We first consider influence analysis in a Bayesian multiple linear regression setting. The data analyzed here comes from the book by Kutner et al. [2004]. This dataset contains 54 test cases. The response variable  $\mathbf{y}$  is the natural logarithm of survival time. There are eight predictor variables: blood-clotting score, prognostic index; enzyme test; liver test; age; gender (binary); moderate alcohol use (binary); and heavy alcohol use (binary). We utilize the following Bayesian model:

$$\begin{aligned} \mathbf{y}_i|\theta, \mathbf{x}_i &\sim f = N(\mathbf{x}_i^T \theta, \sigma^2 \mathbf{I}) \\ \theta &\sim \pi = N(0, 1000\mathbf{I}). \end{aligned}$$

For simplicity, instead of placing a prior on  $\sigma$ , we estimate it from the given data. Because we have chosen a conjugate prior for  $\theta$ , the posterior density is also a Gaussian distribution. Note that if one deletes a case from this data, the resulting posterior distribution is again Gaussian. Furthermore, due to the large differences in predictor scales, we use standardized response and predictor variables to compute the posterior distributions.

In this setup, we are faced with computing the Fisher-Rao distance between two Gaussian posteriors. This requires the computation of a high-dimensional integral, which is not feasible using numerical integration. Thus, we will utilize Monte-Carlo estimation and in particular importance sampling. We note that it is easy to sample from the baseline posterior density; thus we can use it as a natural importance sampling density to estimate the integral given by the Fisher-Rao distance. We rewrite the inner product between the baseline posterior and the posterior under case deletion as follows:

$$\begin{aligned} \langle \sqrt{p_k}, \sqrt{p} \rangle &= \int_{\Theta} \sqrt{p_k(\theta|\mathbf{y}, \mathbf{X})} \sqrt{p(\theta|\mathbf{y}, \mathbf{X})} d\theta \\ &= \int_{\Theta} \frac{\sqrt{p_k(\theta|\mathbf{y}, \mathbf{X})}}{\sqrt{p(\theta|\mathbf{y}, \mathbf{X})}} p(\theta|\mathbf{y}, \mathbf{X}) d\theta. \end{aligned}$$

Thus, our approach is to generate a large sample,  $\{\theta_1, \dots, \theta_N\}$ , from the baseline posterior and then estimate the distance using the following Monte Carlo estimate of  $d_{FR}(p_k, p)$ :

$$(4.1) \quad \hat{d}_{FR}(p_k, p) = \cos^{-1} \left[ \frac{1}{N} \sum_{i=1}^N \frac{\sqrt{p_k(\theta_i|\mathbf{y}, \mathbf{X})}}{\sqrt{p(\theta_i|\mathbf{y}, \mathbf{X})}} \right].$$



Convergence of the estimate  $\hat{d}_{FR}(p_k, p)$  follows via the Ergodic theorem and a continuous mapping argument for the arc cosine.

The results of our analysis are reported in Figure 8. In the left panel, we display the Fisher-Rao distances between the baseline posterior and the posterior under deletion of each case; the middle panel displays the standard Cook’s distance in a frequentist setting; finally, in the right panel we have computed the influence measure proposed in Peng and Dey [1995] based on the Kullback Leibler divergence. We make the following observations: Based on the F statistic, the standard Cook’s distance does not flag any of the observations as influential, even though visually, observation 17 appears influential. Peng and Dey suggest flagging all observations, which yield a “distance” greater than 0.25 as influential under their measure. Thus, using their framework one would consider seven observations as influential, with 17 being highly influential. A similar result can be seen when using the Fisher-Rao distance. Observation 17 is again highly influential (distance greater than 0.7), and there are eight other observations, which can be considered as possibly influential (distance is higher than 0.3).

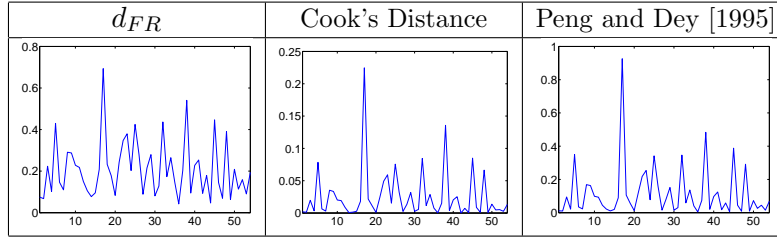


FIGURE 8. Influence analysis in Bayesian multiple linear regression. Each of the 54 observation is listed on the  $x$ -axis with the corresponding measure of influence on the  $y$ -axis.

**4.2. Logistic Regression.** We consider the problem of identifying influential observations in the Bayesian logistic regression setup based on the measure in definition 4. The dataset used here was previously analyzed by Finney [1947] and was studied by Peng and Dey [1995] in influence analysis. There are 39 cases in this data, where the response  $\mathbf{Y}$  is a vector of binary outcomes indicating whether or not vasoconstriction occurred for 39 cases. The two predictor variables, denoted by  $\mathbf{x} = (x_1, x_2)$ , are the volume of air inspired and the rate of air inspired. We consider the logistic model for this data:

$$P(Y_i = 1) = p_i = \frac{e^{\mathbf{x}_i^T \theta}}{1 + e^{\mathbf{x}_i^T \theta}},$$

where  $\theta$  is the unknown vector of coefficients. We assume a multivariate normal distribution for the prior on  $\theta$ , with  $\pi \sim N(\beta_0, \Sigma_0)$ , where  $\beta_0 = \mathbf{1}$  and  $\Sigma_0 = 1000\mathbf{I}$ . Then, the posterior distribution is proportional to

$$f(\mathbf{y}|\mathbf{X}, \theta)\pi(\theta) \propto e^{-0.5(\theta - \theta_0)^T \Sigma_0^{-1}(\theta - \theta_0)} + \sum_{i=1}^{39} (\mathbf{y}_i \mathbf{x}_i^T \theta - \log(1 + e^{\mathbf{x}_i^T \theta})).$$

In this problem  $\theta$  is only three-dimensional and thus we will use numerical integration to obtain the normalizing constant to specify the posterior distribution and to compute the Fisher-Rao distance. The posterior distribution after deletion of case  $k$  can be obtained in

similar fashion. When  $\theta$  is high dimensional one can use Markov Chain Monte Carlo to estimate the integral in the Fisher-Rao distance; we discuss this in detail in the next section. Figure 9 presents the results of our analysis.

We see four clear influential observations (4, 18, 13 and 32 in order of decreasing influence). Observations 4 and 18 appear to have the most severe effect on the posterior distribution of  $\theta$  with resulting distances close to 0.6 or nearly half of the maximal distance on the space of probability densities. The remaining 35 observations yield influence measures lower than 0.2, which we consider as having low influence. We compare our result to that obtained by Peng and Dey [1995]. We refer the reader to their paper for a similar figure as Figure 9 generated under their framework. We note that their influence measure is based on a set of divergences. Thus, it is not symmetric and there is no natural scale on which influence measures can be assessed. The authors suggest a strategy to calibrate the proposed divergence measures but a choice of this calibration is rather arbitrary. Their method flags observations 4 and 18 (in decreasing order of influence) as strong outliers and many other observations as weak outliers. It appears that our approach provides a clearer separation of the influential versus the non-influential observations in this example.

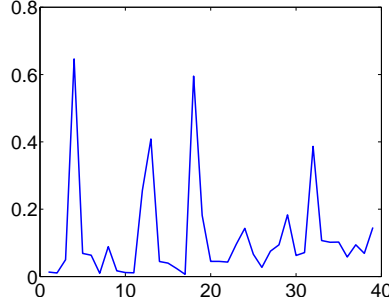


FIGURE 9. Influence analysis in Bayesian logistic regression. Each of the 39 observation is listed on the  $x$ -axis with the corresponding measure of influence ( $d_{FR}$  between baseline posterior and posterior under case deletion) on the  $y$ -axis.

**4.3. MCMC estimate of Fisher-Rao Distance in the case-deletion setting.** In the logistic regression model considered in Section 4.2 for the purposes of influence studies under case-deletion, the posterior density might not be available in closed-form and computing the normalizing constant via a numerical integral might be hard due to a large number of parameters. When this is the case, one can use a Markov Chain Monte Carlo approach to generate samples from the baseline posterior and use the generated samples to evaluate the Fisher-Rao distance between the baseline posterior and the posterior under case deletion. In order to achieve this goal, we must manipulate the expression of the Fisher-Rao distance under the square-root transformation and write it as an expectation with respect to the baseline posterior.

**Proposition 3.** *Suppose  $p_k$  is the posterior density under case deletion and  $p$  is the baseline posterior density; and correspondingly,  $f_k$  and  $f$  are the case deletion and baseline likelihoods with  $\pi$  representing the prior on the parameters. Then,*

$$d_{FR}(p_k, p) = \int_{\Theta} \left[ \frac{f_k(\mathbf{y}|\theta)}{f(\mathbf{y}|\theta)} \right]^{1/2} \left[ \int_{\Theta} \frac{1}{f(y_k|\mathbf{y}_{(k)}, \theta)} p(\theta|\mathbf{y}) d\theta \right]^{-1/2} p(\theta|\mathbf{y}) d\theta.$$

Given this expression and a sample from the posterior density,  $\{\theta_1, \dots, \theta_N\}$ , the Monte Carlo estimate of  $d_{FR}(p_k, p)$  is given by

$$(4.2) \quad \hat{d}_{FR}(p_k, p) = \cos^{-1} \left[ \frac{1}{N} \sum_{i=1}^N a_i b_i \right],$$

where

$$a_i = \left[ \frac{f_k(\mathbf{y}|\theta_i)}{f(\mathbf{y}|\theta)} \right]^{1/2} \quad \text{and} \quad b_i = \left[ \frac{1}{N} \sum_{i=1}^N \frac{1}{f(y_k|\mathbf{y}_{(k)}, \theta_i)} \right]^{-1/2}.$$

It is routine to show that the estimate  $\hat{d}_{FR}(p_k, p)$  is a consistent estimator of  $d_{FR}(p_k, p)$  using the Ergodic theorem.

## 5. CONCLUDING REMARKS

The overarching premise behind this article is on developing statistical procedures based on comparing probability densities. In particular, Bayesian sensitivity analysis provides a fertile background for the construction and deployment of our proposed geometric framework. We provide compelling arguments for the use of the nonparametric Fisher-Rao metric to calculate actual distances between posterior densities as opposed to divergence measures while developing sensitivity measures. This article ought to be appropriately viewed as a first step in developing metric-based sensitivity measures which are naturally calibrated.

The natural next step would be to test the effectiveness of our framework under the nonparametric Bayesian setting. The SRT representation, in principle, would make a seamless transition to that setting from the parametric setup since the manifold of parametric densities is a submanifold of  $\mathcal{P}$  considered here—expressions for geodesic paths and distances remain unaltered. Neighborhoods based on the Kullback-Leibler or the Hellinger divergence are commonly used while assessing posterior consistency. It would be interesting to examine consistency in a geometric neighborhood such as the one considered here; much work remains to be done in this direction.

When posterior densities are unavailable in closed-form, “good” estimators of the geodesic distances are imperative. Excepting the setting of influence analysis under case-deletion, this has not been explored in this article and is of some importance. Methods of incorporating the calculation of the geodesics into existing MCMC procedures would be greatly beneficial. However, under the parametric setting when the unknown parameter vector is of small dimension—similar to the settings considered in this article—the geodesic distances can be calculated with a fair degree of accuracy.

## 6. ACKNOWLEDGMENTS

We would like to thank Steve Maceachern, Dipak Dey and Anuj Srivastava for the useful discussions and suggestions.

## APPENDIX

### Proof of Theorem 1:

*Proof.* Let  $r$  be a small positive scalar and  $\delta p \in T_p(\mathcal{P})$ . We begin by computing the differential of the mapping  $\phi, \phi_* : T_p(\mathcal{P}) \rightarrow T_{\phi(p)}(\Psi)$

$$\phi_*(\delta p) = \left. \frac{d}{dr} \phi(p + r\delta p) \right|_{r=0} = \left. \frac{d}{dr} \sqrt{p + r\delta p} \right|_{r=0} = \left. \frac{\delta p}{2\sqrt{p + r\delta p}} \right|_{r=0} = \frac{\delta p}{2\sqrt{p}}$$

Plugging this expression into the standard  $\mathbb{L}^2$  metric, for two tangent vectors  $\delta p_1, \delta p_2 \in T_p(\mathcal{P})$ , we obtain the following:

$$\langle \phi_*(\delta p_1), \phi_*(\delta p_2) \rangle = \left\langle \frac{\delta p_1}{2\sqrt{p}}, \frac{\delta p_2}{2\sqrt{p}} \right\rangle = \frac{1}{4} \int_{\mathbb{R}} \delta p_1(x) \delta p_2(x) (1/p(x)) dx = \frac{1}{4} \langle \delta p_1, \delta p_2 \rangle_p.$$

□

### Proof of Theorem 2:

*Proof.* Let  $f(x|\theta)$  be the likelihood function,  $\pi_0(\theta)$  be the baseline prior, and  $g \in \mathcal{G}$  represent a contamination density. We begin by focusing on the linear contamination class,  $\Gamma$ . We can write a perturbation of the baseline prior as  $\delta_l \pi_0 = g - \pi_0$ , and the  $\epsilon$ -contaminated prior is given by  $\pi_l = \pi_0 + \epsilon \delta_l \pi_0$ . Then, the contaminated joint density is given by  $p_l(z, \theta) = f(z|\theta) \pi_l(\theta)$ . Next, we utilize the  $\mathbb{L}^2$  metric on the space of SRT representations of densities. We compute the perturbation vector on this space as follows:

$$\begin{aligned} v_l(z, \theta) &= \frac{d}{d\epsilon} \sqrt{f(z|\theta) \pi_l(\theta)}|_{\epsilon=0} \\ &= \frac{d}{d\epsilon} \sqrt{f(z|\theta) \pi_0(\theta) + \epsilon f(z|\theta) \delta_l \pi_0}|_{\epsilon=0} \\ &= \frac{f(z|\theta) \delta_l \pi_0}{2\sqrt{f(z|\theta) \pi_0(\theta)}} \\ &= \sqrt{f(z|\theta)} \frac{(g - \pi_0)}{2\sqrt{\pi_0}}. \end{aligned}$$

Given two perturbations of the baseline prior  $\delta_l^1 \pi_0, \delta_l^2 \pi_0$ , we compute the corresponding perturbations of the joint density under the SRT representation and derive the corresponding inner product on that space:

$$\begin{aligned} \langle v_l^1, v_l^2 \rangle &= \int_{\Theta} \int_{\mathbb{R}} \frac{g_1(\theta) - \pi_0(\theta)}{2\sqrt{\pi_0(\theta)}} \frac{g_2(\theta) - \pi_0(\theta)}{2\sqrt{\pi_0(\theta)}} f(z|\theta) dz d\theta \\ &= \frac{1}{4} \int_{\Theta} \frac{(g_1(\theta) - \pi_0(\theta))(g_2(\theta) - \pi_0(\theta))}{\pi_0(\theta)} d\theta, \end{aligned}$$

which is independent of the sampling distribution. This result was previously derived in Zhu et al. [2011] for the parametric version of the Fisher Rao metric under the log transformation of probability density functions. Next, we focus on the geometric contamination class,  $\tilde{\Gamma}$ . We write a perturbation of the baseline prior (using the SRT representation) as  $\delta_{gm} \sqrt{\pi_0} = \exp_{\sqrt{\pi_0}}^{-1}(\sqrt{g})$ . Then the SRT representation of the contaminated prior is given by  $\sqrt{\pi_{gm}} = \exp_{\sqrt{\pi_0}}(\epsilon \delta_{gm} \sqrt{\pi_0})$ . The exponential and inverse exponential maps that are used here were defined in Section 2.3. The SRT representation of the contaminated joint density is given by  $\sqrt{p_{gm}(z, \theta)} = \sqrt{f(z|\theta) \pi_{gm}(\theta)}$ . We compute the perturbation vector on the space of SRT

representations of joint densities as follows:

$$\begin{aligned}
v_{gm}(z, \theta) &= \frac{d}{d\epsilon} \sqrt{f(z|\theta)\pi_{gm}(\theta)}|_{\epsilon=0} \\
&= \frac{d}{d\epsilon} \sqrt{f(z|\theta)} \exp_{\sqrt{\pi_0}}(\epsilon \delta_{gm} \sqrt{\pi_0})|_{\epsilon=0} \\
&= \frac{d}{d\epsilon} \sqrt{f(z|\theta)} (\cos(\epsilon \|\delta_{gm} \sqrt{\pi_0}\|) \sqrt{\pi_0} + \sin(\epsilon \|\delta_{gm} \sqrt{\pi_0}\|) \frac{\delta_{gm} \sqrt{\pi_0}}{\|\delta_{gm} \sqrt{\pi_0}\|})|_{\epsilon=0} \\
&= \sqrt{f(z|\theta)} (-\sin(\epsilon \|\delta_{gm} \sqrt{\pi_0}\|) \sqrt{\pi_0} \|\delta_{gm} \sqrt{\pi_0}\| + \cos(\epsilon \|\delta_{gm} \sqrt{\pi_0}\|) \delta_{gm} \sqrt{\pi_0})|_{\epsilon=0} \\
&= \sqrt{f(z|\theta)} \delta_{gm} \sqrt{\pi_0}
\end{aligned}$$

Given two geometric perturbations of the baseline prior  $\delta_{gm}^1 \sqrt{\pi_0}$ ,  $\delta_{gm}^2 \sqrt{\pi_0}$ , we compute the corresponding perturbations of the joint density under the SRT representation and derive the corresponding inner product on that space:

$$\begin{aligned}
\langle v_{gm}^1, v_{gm}^2 \rangle &= \int_{\Theta} \int_{\mathbb{R}} \delta_{gm}^1 \sqrt{\pi_0}(\theta) \delta_{gm}^2 \sqrt{\pi_0}(\theta) f(z|\theta) dz d\theta \\
&= \int_{\Theta} \delta_{gm}^1 \sqrt{\pi_0}(\theta) \delta_{gm}^2 \sqrt{\pi_0}(\theta) d\theta,
\end{aligned}$$

which is again independent of the sampling distribution. Equation 6 further suggests that if the sampling distribution is fixed and the geometric perturbation model is used, the metric on the space of joint densities is the same as that on the space of priors. Intuitively, one would expect this to be the case and thus this is an attractive property of our framework.  $\square$

#### *Proofs of results concerning local perturbations*

For notational convenience we use the same notation as given in the main text and set  $e(\theta)$  to denote  $\exp_{\sqrt{\pi_0}}(\epsilon v)(\theta)$ ,  $de(\theta)$  to denote  $\frac{d}{d\epsilon} \exp_{\sqrt{\pi_0}}(\epsilon v)(\theta)$ , and  $d^2e(\theta)$  to denote  $\frac{d^2}{d\epsilon^2} \exp_{\sqrt{\pi_0}}(\epsilon v)(\theta)$ . We use the following results:

$$\begin{aligned}
\exp_{\sqrt{\pi_0}}(\epsilon v_g)|_{\epsilon=0} &= \cos(\epsilon \|v_g\|) \sqrt{\pi_0} + \sin(\epsilon \|v_g\|) \frac{v_g}{\|v_g\|}|_{\epsilon=0} = \sqrt{\pi_0}, \\
\frac{d}{d\epsilon} \exp_{\sqrt{\pi_0}}(\epsilon v_g)|_{\epsilon=0} &= -\sin(\epsilon \|v_g\|) \sqrt{\pi_0} \|v_g\| + \cos(\epsilon \|v_g\|) v_g|_{\epsilon=0} = v_g, \\
\frac{d^2}{d\epsilon^2} \exp_{\sqrt{\pi_0}}(\epsilon v_g)|_{\epsilon=0} &= -\cos(\epsilon \|v_g\|) \sqrt{\pi_0} \|v_g\|^2 - \sin(\epsilon \|v_g\|) v_g \|v_g\| |_{\epsilon=0} = -\sqrt{\pi_0} \|v_g\|^2.
\end{aligned}$$

#### **Proof of Proposition 1:**

$$\begin{aligned}
\frac{d}{d\epsilon} F_{\pi_0}(v_g)|_{\epsilon=0} &= \frac{d}{d\epsilon} \frac{m(x|\pi_0)}{m(x|\epsilon g)} \Big|_{\epsilon=0} \\
&= -2m(x|\pi_0) \frac{\int_{\Theta} f(x|\theta) de(\theta) e(\theta) d\theta}{(\int_{\Theta} f(x|\theta) e(\theta)^2 d\theta)^2} \Big|_{\epsilon=0} \\
&= -2 \frac{m(x|v_g)}{m(x|\pi_0)}.
\end{aligned}$$

#### **Proof of Proposition 2:**

$$\begin{aligned}
\left. \frac{d}{d\epsilon} F_{\pi_0}(v_g) \right|_{\epsilon=0} &= \left. \frac{d}{d\epsilon} \int_{\Theta} h(\theta) \frac{f(x|\theta)e(\theta)^2}{\int_{\Theta} f(x|\theta)e(\theta)^2 d\theta} d\theta \right|_{\epsilon=0} \\
&= 2 \int_{\Theta} h(\theta) \frac{f(x|\theta)e(\theta)de(\theta) \int_{\Theta} f(x|\theta)e(\theta)^2 d\theta - f(x|\theta)e(\theta)^2 \int_{\Theta} f(x|\theta)e(\theta)de(\theta)d\theta}{(\int_{\Theta} f(x|\theta)e(\theta)^2 d\theta)^2} d\theta \Big|_{\epsilon=0} \\
&= 2 \int_{\Theta} h(\theta) \frac{f(x|\theta)\sqrt{\pi_0(\theta)}v(\theta)}{m(x|\pi_0)} d\theta - 2 \frac{m(x|v_g)}{m(x|\pi_0)} \int_{\Theta} h(\theta)p_0(\theta|x)d\theta.
\end{aligned}$$

**Proof of Theorem 3:**

We note that since we are dealing with infinitesimal quantities, we make a simplification using the local Euclidean structure of  $\Psi$ , by approximating the arc length distance using a chord length distance, which locally are essentially the same (see equation 2.9 Kass [1989]). Indeed every manifold is locally Euclidean and we exploit this property in the proof in the sense that the geodesic distance based on the nonparametric Fisher-Rao metric is well approximated by the  $\mathbb{L}^2$  distance

$$\left\| \sqrt{p_0} - \sqrt{\frac{f \exp_{\sqrt{\pi_0}}(\epsilon v_g)^2}{m(x|\epsilon g)}} \right\|^2.$$

Therefore,

$$\begin{aligned}
\left. \frac{d}{d\epsilon} F_{\pi_0}(v_g) \right|_{\epsilon=0} &= \left. \frac{d}{d\epsilon} \left\| \sqrt{p_0} - \sqrt{\frac{f \exp_{\sqrt{\pi_0}}(\epsilon v_g)^2}{m(x|\epsilon g)}} \right\|^2 \right|_{\epsilon=0} \\
&= \frac{d}{d\epsilon} \int_{\Theta} \left[ \sqrt{p_0(\theta|x)} - \sqrt{\frac{f(x|\theta)e(\theta)^2}{m(x|\epsilon g)}} \right]^2 d\theta \Big|_{\epsilon=0}
\end{aligned}$$

We now take the derivative inside the integral leading to an ungainly looking expression:

$$\begin{aligned}
\left. \frac{d}{d\epsilon} F_{\pi_0}(v_g) \right|_{\epsilon=0} &= \int_{\Theta} \left[ \sqrt{p_0(\theta|x)} - \sqrt{\frac{f(x|\theta)e(\theta)^2}{m(x|\epsilon g)}} \right] \sqrt{\frac{m(x|\epsilon g)}{f(x|\theta)e(\theta)^2}} \\
&\quad \left[ \frac{f(x|\theta)e(\theta)de(\theta) \int_{\Theta} f(x|\theta)e(\theta)^2 d\theta - f(x|\theta)e(\theta)^2 \int_{\Theta} f(x|\theta)e(\theta)de(\theta)d\theta}{m(x|\epsilon g)^2} \right] d\theta \Big|_{\epsilon=0}.
\end{aligned}$$

Now, for convenience, let

$$T(\theta) = \left[ \frac{f(x|\theta)e(\theta)de(\theta) \int_{\Theta} f(x|\theta)e(\theta)^2 d\theta - f(x|\theta)e(\theta)^2 \int_{\Theta} f(x|\theta)e(\theta)de(\theta)d\theta}{m(x|\epsilon g)^2} \right]$$

Then, we have a simplified-looking expression

$$\begin{aligned}
\left. \frac{d}{d\epsilon} F_{\pi_0}(v_g) \right|_{\epsilon=0} &= \int_{\Theta} \left[ \sqrt{p_0(\theta|x)} \sqrt{\frac{m(x|\epsilon g)}{f(x|\theta)e(\theta)^2}} - 1 \right] T(\theta) d\theta \Big|_{\epsilon=0} \\
&= \int_{\Theta} \left[ \sqrt{\frac{p_0(\theta|x)}{p_0(\theta|x)}} - 1 \right] \left[ \frac{f(x|\theta)\sqrt{\pi_0(\theta)}v(\theta) - p_0(\theta|x)m(x|v_g)}{m(x|\pi_0)} \right] d\theta \\
&= 0,
\end{aligned}$$

as expected since the distance is minimized at 0. This compels us to consider the second derivative to obtain a finer measure. The second derivative with respect to  $\epsilon$  is

$$\begin{aligned}
\left. \frac{d^2}{d\epsilon^2} F_{\pi_0}(v_g) \right|_{\epsilon=0} &= \frac{d^2}{d\epsilon^2} \int_{\Theta} \left[ \sqrt{p_0(\theta|x)} - \sqrt{\frac{f(x|\theta)e(\theta)^2}{m(x|\epsilon g)}} \right]^2 d\theta \Big|_{\epsilon=0} \\
&= \int_{\Theta} \sqrt{\frac{m_0(x|\pi_0)e(\theta)^2}{m(x|\epsilon g)\pi_0(\theta)}} \left[ \frac{m(x|\pi_0)\pi_0(\theta)e(\theta)^2 \int_{\Theta} f(x|\theta)e(\theta)de(\theta)d\theta}{m(x|\pi_0)^2e(\theta)^4} \right. \\
&\quad \left. - \frac{m(x|\pi_0)\pi_0(\theta)e(\theta)de(\theta)m(x|\epsilon g)}{m(x|\pi_0)^2e(\theta)^4} \right] T(\theta)d\theta + \int_{\Theta} \left[ \sqrt{\frac{m(x|\epsilon g)\pi_0(\theta)}{m_0(x|\pi_0)e(\theta)^2}} - 1 \right] dT(\theta)d\theta \Big|_{\epsilon=0} \\
&= \int_{\Theta} \left[ \sqrt{\frac{m(x|\pi_0)\pi_0(\theta)}{m(x|\pi_0)\pi_0(\theta)}} \frac{\pi_0(\theta)m(x|v_g) - \sqrt{\pi_0(\theta)}v(\theta)m(x|\pi_0)}{m(x|\pi_0)\pi_0(\theta)} \right. \\
&\quad \left. \left[ \frac{f(x|\theta)\sqrt{\pi_0(\theta)}v(\theta)m(x|\pi_0) - f(x|\theta)\pi_0(\theta)m(x|v_g)}{m(x|\pi_0)^2} \right] d\theta \right. \\
&\quad \left. + \int_{\Theta} \left[ \sqrt{\frac{m(x|\pi_0)\pi_0(\theta)}{m(x|\pi_0)\pi_0(\theta)}} - 1 \right] dT(\theta)d\theta \right] \Big|_{\epsilon=0} \\
&= 4 \frac{m(x|v_g)}{m(x|\pi_0)} \int_{\Theta} \frac{v(\theta)}{\sqrt{\pi_0(\theta)}} p_0(\theta|x) d\theta - 2 \int_{\Theta} \frac{v(\theta)^2}{\pi_0(\theta)} p_0(\theta|x) d\theta - 2 \frac{m(x|v_g)^2}{m(x|\pi_0)^2}.
\end{aligned}$$

**Proof of Proposition 3:**

*Proof.* A key observation here is that under the case deletion setup, the prior on the parameters does not change. Now, we have the following:

$$\begin{aligned}
d_{FR}(p_k, p) &= \int_{\Theta} \sqrt{p_k(\theta|y)} \sqrt{p(\theta|y)} d\theta \\
&= \int_{\Theta} \sqrt{p_k(\theta|y)} \frac{1}{\sqrt{p(\theta|y)}} p(\theta|y) d\theta
\end{aligned}$$

Now, substituting the expression for the posteriors based on the likelihoods and the prior, we obtain

$$\begin{aligned}
d_{FR}(p_k, p) &= \int_{\Theta} \sqrt{\frac{f_k(y|\theta)\pi(\theta)}{\int_{\Theta} f_k(y|\theta)\pi(\theta)d\theta}} \sqrt{\frac{\int_{\Theta} f(y|\theta)\pi(\theta)d\theta}{f(y|\theta)\pi(\theta)}} p(\theta|y) d\theta \\
&= \int_{\Theta} \sqrt{\frac{f_k(y|\theta)}{f(y|\theta)}} \sqrt{\frac{\int_{\Theta} f(y|\theta)\pi(\theta)d\theta}{\int_{\Theta} f_k(y|\theta)\pi(\theta)d\theta}} p(\theta|y) d\theta \\
&= \int_{\Theta} \left[ \frac{f_k(y|\theta)}{f(y|\theta)} \right]^{1/2} \left[ \int_{\Theta} \frac{1}{f(y_k|y_{(k)}, \theta)} p(\theta|y) d\theta \right]^{-1/2} p(\theta|y) d\theta.
\end{aligned}$$

□

REFERENCES

- S. Amari. *Differential Geometric Methods in Statistics*. Lecture Notes in Statistics, Vol. 28. Springer, 1985.

- S I Amari, O E Barndorff-Nielsen, R E Kass, S L Lauritzen, and C R Rao. *Differential Geometry in Statistical Inference*. Institute of Mathematical Statistics, Lecture Notes-Monograph Series, 1987.
- J O Berger. Robust bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25:303–328, 1990.
- J O Berger. An overview of robust bayesian analysis. *TEST*, 3:5–58, 1994.
- J O Berger and L M Berliner. Robust bayes and empirical bayes analysis with  $\epsilon$ -contaminated prior. *Annals of Statistics*, 14:461–486, 1986.
- A. Bhattacharya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of Calcutta Mathematical Society*, 35:99–109, 1943.
- B P Carlin and N G Polson. An expected utility approach to influence diagnostics. *Journal of the American Statistical Association*, 86(1):1013–1021, 1991.
- K M Carter, R Raich, W G Fimm, and A O Hero. Fine: fisher information nonparametric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2093–2098, 2009.
- N. N. Čencov. *Statistical Decision Rules and Optimal Inferences*, volume 53 of *Translations of Mathematical Monographs*. AMS, 1982.
- I Csiszar. *Information Measures: A Critical Survey, Trans. 7th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*,. Reidel, Dordrecht, 1974.
- D K Dey and L R Birmiwai. Robust bayesian analysis using divergence measures. *Statistics and Probability Letters*, 20(1):287–294, 1994.
- D.J. Finney. The estimation from individual records of the relationship between dose and quantal response. *Biometrika*, 34:320–334, 1947.
- I Guttman and D Pena. A bayesian look at diagnostics in the univariate linear model. *Statistica Sinica*, 3(1):367–390, 1993.
- D R Insua and F Ruggeri. *Lecture Notes in Statistics: Robust Bayesian Analysis*. Springer, New York., 2000.
- R Kass. The geometry of asymptotic inference. *Statistical Science*, 3(1):188–219, 1989.
- M.H. Kutner, C.J. Nachtsheim, and J. Neter. *Applied Linear Regression Models*. McGraw-Hill/Irwin, fourth international edition, 2004.
- S. Lang. *Fundamentals of Differential Geometry*. Springer, 1999.
- E Moreno. Global bayesian robustness for some classes of prior distributions. In D R Insua and F Ruggeri, editors, *Lecture Notes in Statistics: Robust Bayesian Analysis*, pages 45–70. Springer, 2000.
- F. Peng and D.K. Dey. Bayesian analysis of outlier problems using divergence measures. *The Canadian Journal of Statistics*, 23(2):199–213, 1995.
- C. R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.
- A. Srivastava, I. Jermyn, and S.H. Joshi. Riemannian analysis of probability density functions with applications in vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- A. Srivastava, E. Klassen, S.H. Joshi, and I.H. Jermyn. Shape analysis of elastic curves in euclidean spaces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 33(7):1415–1428, 2011.
- A. Srivastava, W. Wu, S. Kurtek, E. Klassen, and J. S. Marron. Statistical analysis and modeling of elastic functions. *ArXiv e-prints*, 2011.



P. W. Vos and R. E. Kass. *Geometrical Foundations of Asymptotic Inference*. Wiley-Interscience, 1997.

H Zhu, J G Ibrahim, and Niansheng Teng. Bayesian influence analysis: A geometric approach. *Biometrika*, 98(2):307–323, 2011.

DEPARTMENT OF STATISTICS; 404 COCKINS HALL; 1958 NEIL AVENUE; THE OHIO STATE UNIVERSITY; COLUMBUS, OH 43210.

*E-mail address:* `kurtek.1@stat.osu.edu`

DEPARTMENT OF STATISTICS; 404 COCKINS HALL; 1958 NEIL AVENUE; THE OHIO STATE UNIVERSITY; COLUMBUS, OH 43210.

*E-mail address:* `bharath.4@osu.edu`